HOMIE: Human-object Centric Video Personalization via Multimodal Intelligent Enhancement

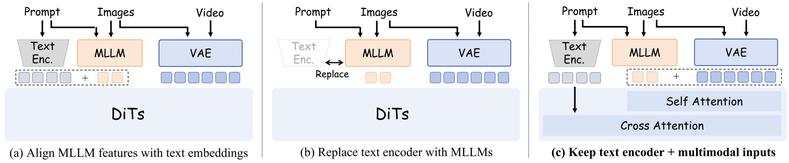
Yiyang Cai*, Nan Chen*, Rongchang Xie, Junwen Pan, Chunyang Jiang, Cheng Chen, Chang Liu , Zhenbang Sun, Wei Xue, Wenhan Luo†, Yike Guo

Contact Info

Video Demos and Codes here

Introduction

Human-object centric video personalization (HOCVP) **generates videos with the user-defined human and object**. It enables applications such as product demonstrations, advertising videos, virtual avatars, and brand content creation.



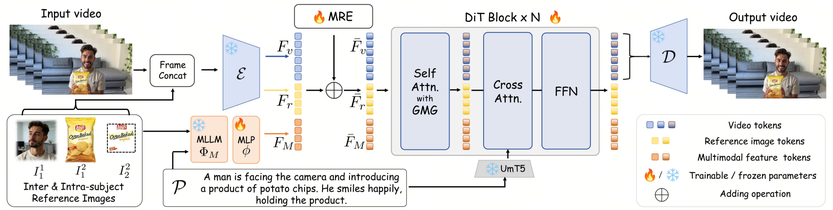
(a) Align MLLM features with text embeddings (b) Replace text encoder with MLLMs **(c) Keep text encoder + multimodal inputs**

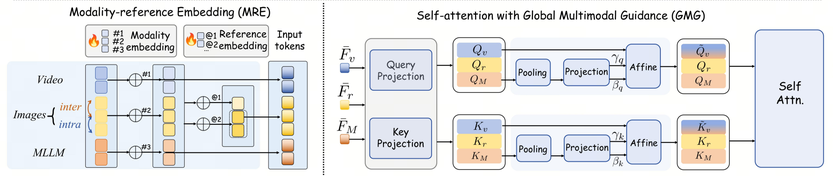
Better understanding drives better generation. However, previous MLLM integration strategies are suboptimal because: 1. Aligning MLLM features with text features can weaken text controllability; 2. Replacing the text encoder with an MLLM makes training more difficult. (See strategy a&b above)



HOMIE is a novel HOCVP framework, which proposes:
1. A new MLLM integration strategy that preserves original text control while enhancing multimodal understanding.
2. A reorganized reference input structure that supports both inter-subject and intra-subject video personalization.

Methodology



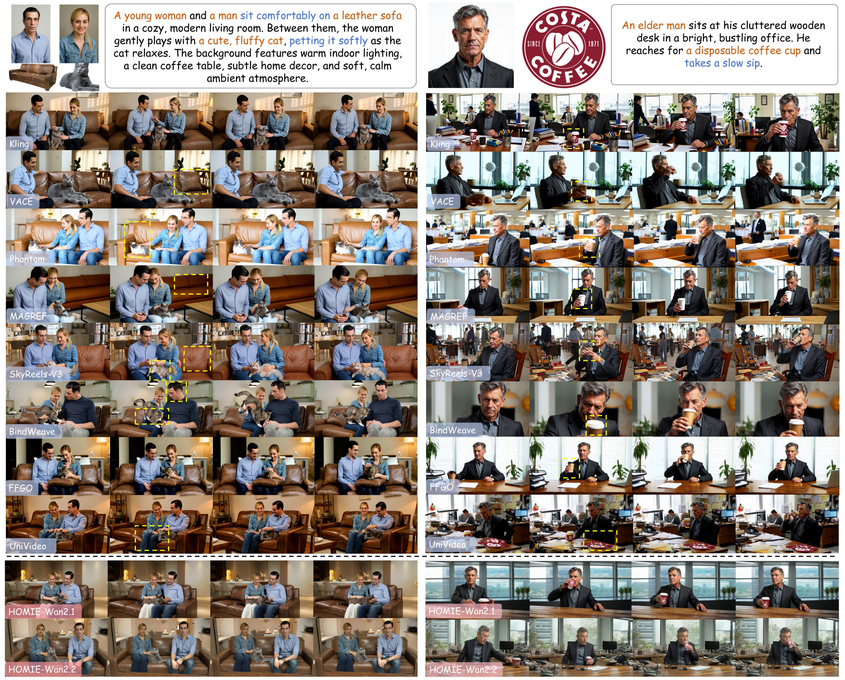


HOMIE has two core designs:
GMG (Global Multimodal Guidance) injects MLLM-based global understanding into video generation
MRE (Modality-Reference Embedding) enables the model to distinguish text, video, and reference inputs, as well as different subjects and multiple references of the same subject.

Quantitative Results

Method	Video Quality		Text Following	Subject Consistency			
	AES↑	MS↑	GMEScore↑	Face-Sim↑	DINO-I↑	Obj-Sim ↑	OCR Acc.↑
Kling 1.6 [29]	**0.558**	**0.997**	0.531	0.678	0.492	0.764	0.272
VACE [28]	0.517	0.994	0.685	0.650	0.539	0.828	0.371
MAGREF [15]	0.499	0.971	0.689	0.589	0.429	0.754	0.135
SkyReels-A2 [16]	0.493	0.946	0.652	0.599	0.524	0.803	0.288
SkyReels-V3 [32]	0.523	0.971	0.656	0.739	0.526	0.847	0.326
Phantom [43]	0.505	0.991	0.677	**0.801**	0.519	0.774	0.302
HuMo [9]	0.468	0.988	0.649	0.749	0.463	0.734	0.304
BindWeave [35]	0.471	0.970	0.610	0.635	0.399	0.557	0.111
FFGO [7]	0.413	0.939	0.673	0.580	0.422	0.739	0.235
UniVideo [35]	0.463	0.995	0.684	0.772	0.493	0.872	0.325
VINO [8]	0.506	0.969	0.694	0.768	0.509	0.868	0.331
Ours (Wan2.1-14B)	0.508	0.994	0.691	0.786	**0.543**	**0.891**	**0.452**
Ours (Wan2.2-14B)	0.525	**0.997**	**0.696**	0.760	0.537	0.877	0.431

Quanlitative Results



Ablation Studies



(a) Ablation Study of Different MLLM Integration Strategies



(b) Ablation Study of the MRE Module